CLEF-2005 CL-SR at Maryland: Document and Query Expansion using Side Collections and Thesauri

Jianqiang Wang and Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
(wangjq,oard)@glue.umd.edu

Abstract

This paper reports results for the University of Maryland's participation in CLEF-2005 Cross-Language Speech Retrieval track. Techniques that were tried include: (1) document expansion with manually created metadata (thesaurus keywords and segment summaries) from a large side collection, (2) query refinement with pseudo-relevance feedback, (3) keyword expansion with thesaurus synonyms, and (4) cross-language speech retrieval using translation knowledge obtained from the statistics of a large parallel corpus. The results show that document expansion and query expansion using blind relevance feedback were effective, although optimal parameter choices differed somewhat between the training and evaluation sets. Document expansion in which manually assigned keywords were augmented with thesaurus synonyms yielded marginal gains on the training set, but no improvement on the evaluation set. Cross-language retrieval with French queries yielded 79% of monolingual mean average precision when searching manually assigned metadata despite a substantial domain mismatch between the parallel corpus and the retrieval task. Detailed failure analysis indicates that speech recognition errors for named entities were an important factor that substantially degraded retrieval effectiveness.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Speech Retrieval, Document Expansion, Query Expansion, Blind Relevance Feedback

1 Introduction

Automated techniques for speech retrieval seek to provide users with access to spoken content. Although manual transcription and manual cataloging of spoken word collections are widely used, manual transcription suffers from limited scalability and recording-level manual cataloging suffers from limited specificity. The most widely adopted approaches to fully automated content-based speech retrieval rely on the combination of two critical techniques: automatic speech recognition

(ASR) and information retrieval (IR). An ASR engine is first used to transcribe digitized audio into text, and text retrieval techniques can then be applied to accomplish the task. However, since ASR is an imperfect process, often there are spoken words that are not recognized correctly. That will lead to word mismatch in the retrieval step. Therefore, improving ASR accuracy (i.e., decreasing the ASR word error rate (WER)) can improve retrieval effectiveness [3]. This doesn't mean perfect ASR is a necessity, however. Early experiments with speech retrieval for broadcast news in the TREC Spoken Document Retrieval (SDR) track showed shown that modern ranked retrieval techniques are fairly robust in the presence of speech recognition errors. For example, Word Error Rates (WER) as high as 40% were observed to degrade retrieval effectiveness by less than 10% [1]. Routinely achieving that level of accuracy for broadcast news is now well within the state of the art.

The challenge of automated access to spoken content is, however, far from completely solved because broadcast news represents only a small portion of the variety of spoken content that information users may be interested in. Examples of other types of spoken word collections include recordings of calls to help desks, political speeches, meetings, and lectures. This year's CLEF Cross-Language Speech Retrieval (CL-SR) track chose oral history interviews. The collection contains automatically transcribed text from 291 manually segmented interviews with Holocaust survivors, witnesses and rescuers, together with manually generated thesaurus keywords and segment summaries and automatically produced keywords. This offers an excellent opportunity to study the application of techniques that have proven to be successful for searching broadcast news to a different domain, while providing opportunities to explore additional issues that are not easily studied in news genre.

In this study, we first wanted to re-examine how speech recognition errors affect IR effectiveness. The ASR text was produced by an IBM speech recognition system that was specifically trained for this specific collection. The WER for that system is about 29%, which the reported results with broadcast news would suggest should be adequate. An initial study that we conducted in 2004 using a smaller number of topics and a less well vetted set of relevance judgments indicated, however, that retrieval effectiveness when searching that collection using ASR results was substantially below what we could obtain when using either manually transcribed text or manually assigned metadata [5]. The improved ASR accuracy and the larger number of topics in the CLEF-2005 CL-SR collection permits a more thorough exploration of the reasons for that effect. Second, query and document expansion using blind relevance feedback are known to improve retrieval effectiveness when applied to broadcast news but we are not aware of similar experiments with any source of spontaneous speech. The availability of a training/evaluation split among the CLEF-2005 CL-SR topics makes it possible to explore this question in a principled manner. Expansion techniques require a large collection of text with similar topical content; we used manually generated thesaurus keywords and summaries from other interviews for this purpose. Finally, the availability of topics in languages other than English facilitates cross-language speech retrieval experiments. Translation probabilities obtained by training statistical models with parallel text have proven to be very useful for ranked retrieval of newswire stories (e.g., [12, 2]), and we were interested in assessing the effectiveness of those techniques when applied to a different domain. We chose to focus on manually assigned metadata for these initial experiments in order to avoid introducing confounding effects from ASR errors.

The remainder of this paper is organized as follows. In the next section, we describe the techniques that we applied. Section 3 then presents mean average precision results for our five official submissions and additional experiments that we scored locally using both the training and the evaluation collections. Section 4 augments those results with an initial query-by-query analysis of the effect of ASR errors. The paper then concludes with a few remarks on our future plans.

2 Techniques

In this section we describe the techniques that we used in our experiments.

2.1 Document expansion using blind relevance feedback

There are generally two types of errors that an ASR system can produce: (1) failure to recognize some spoken words (2) introduction of spurious words. These problems often occur together: because ASR systems seek to map sounds to words, recognition errors generally lead to mapping the associated sounds to spurious words. Missing words reduce *word-recall* (proportion of spoken words that are recognized) while adding words reduce *word precision* (proportion of recognized words that were spoken). Singhal et al argue that IR would benefit from high word-recall, and that it would be less influenced by poor word precision [9]. Our observations from a few randomly selected samples of audio and ASR text from CLEF-CL-SR collection seem to confirm their argument. In most cases, the spurious words that were incorrectly added by the speech recognizer were common words, which would be expected to have a relatively small impact on retrieval effectiveness. Many of the missing spoken words, by contrast, were content-bearing words such as named entities, which could be crucial for accurate retrieval. Singhal et al proposed an approach that they called *document expansion* that enriched each document in the collection with additional words that "could have been there." Specifically, the one-best speech recognition hypothesis for each "speech document" was used as a query to retrieve related documents from a side collection of newswire text. A few highly selective terms were then selected from the most highly ranked newswire text documents and added to the original speech documents. Those speech documents were then re-indexed so that subsequent searches could match on the words that were added. In their TREC paper and their later SIGIR paper, they described different ways of tuning the parameters of this technique (e.g., how many highly ranked documents and how many of the most selective terms) and how those choices influenced the effectiveness of document expansion [9, 10]. They found that document expansion yielded substantial improvements in retrieval effectiveness, although the lack of a training/evaluation split in the TREC SDR collection that they used precluded exploration of the stability of the optimal parameters across topic sets.

Applying document expansion to the CLEF-2005 CL-SR test collection required that we identify a source of documents that can be used as a basis for expansion. One obvious choice would be texts about the Holocaust, but assembling a large collection of such texts proved to be more difficult than finding large collections of news stories. Fortunately, another source of clean text was close at hand. The interviews in the CLEF-2005 CL-SR test collection were selected from 4,780 English interviews that were manually indexed at the Survivors the Shoah Visual History Foundation in a similar manner. A subset of 403 of those interviews are reserved for use in the CLEF-2005 test collection (although ASR availability results in only 291 of those interviews having been included in the 2005 release of the test collection). This results in 4,377 interviews being usable for fair expansion experiments. The manual indexing process generated several types of metadata, including segment boundaries, thesaurus keywords and segment summaries. After excluding short segments in which a displayed physical object was the primary referent (this fact is indicated by a manually assigned thesaurus term), we obtained 168,584 segments that could be used as a basis for document expansion. On average, each segment summary has about 30 words, with a minimum of 0 (61 segments somehow don't have any summary) and a maximum of 385. In this expansion collection, there are on average about 4 thesaurus terms per segment, with a minimum of 1 and a maximum of 30. Thesaurus terms are usually multi-word expressions, typically containing 2-4 words. We formed 168,584 documents for the expansion collection, each with an average of 48 words (minimum 2, maximum 417), by combining the summaries and the preferred version of each thesaurus term. Synonyms and other thesaurus relationships were not used.[1]

There are some obvious advantages of using this side collection to perform document expansion on ASR text. From our 2004 experiments and our experience with the CLEF-2005 CL-SR training collection we have clearly seen that manual metadata yields better retrieval effectiveness than the

[1] Unfortunately, this expansion collection could not be distributed to other CLEF-2005 participants because special data protection agreements that are beyond the scope of the standard CLEF agreement would be required. Teams interested in obtaining this collection for research use in 2006 should contact us early to discuss the technical and legal issues involved.

presently available ASR results. If we could enrich the ASR text with well chosen terms from manual metadata, we would therefore reasonably expect to improve retrieval effectiveness. This is important an important question since it would open the door to reusing manual metadata that has already been produced to improve the retrieval of additional interviews without the expense of additional manual indexing (e.g., there are more than 20,000 additional English interviews for which no manually created segment summaries will be created). Another advantage to using this collection is that manually created metadata from the same segments was used to train one of the text classifiers used to produce the automatic keywords that were provided with the CLEF-2005 CL-SR test collection. Results with blind relevance feedback experiments for document expansion can therefore help to inform the design of k-Nearest-Neighbor techniques for automatically assigning thesaurus terms.

The present structure of the test collection imposed some limitations on our document expansion experiments. First, word lattices that encoded alternate hypotheses from the ASR experiments were not available, so it was not possible to limit the expansion words to those that appear somewhere in the word lattice. Singhal et al had found that such a restriction could be useful [9]. Second, the ASR text for each segment contains an average of 503 words. Query processing time grows roughly linearly with the length of the query, so it would be computationally impractical to use every word produced by ASR as a query, even for this relatively small 8,104-segment test collection. We therefore tried two techniques for ranking terms for query selection: (1) Robertson Sparck Jones offer weights and (2) Okapi BM 25 weights [7]. Experiments with the training set indicated that Okapi weights were the better choice in this case.

Specifically, our implementation of document expansion works as follows:

Formulating document-queries. ASR tokens from each document in the test collection are ranked with their Okapi term weight, and the top-20 and top-40 words are selected to formulate two sets of queries. We created two sets instead of one in order to see how the number of words selected to represent each document affects document expansion results. The Okapi weight we used is:

$$w = [\log \frac{(N - df + 0.5)}{(df + 0.5)}][\frac{(2.2 * tf)}{(0.3 + 0.9 * \frac{dl}{avdl} + tf)}] \tag{1}$$

where

- N is the size of the test collection.
- df is the number of documents that contain the term.
- tf is the frequency of the term in the document.
- dl is the length of the document.
- $avdl$ is the average document length of the test collection.

Searching the side collection. We used these queries to search the side collection for the most closely related segments based on lexical overlap with the summary and thesaurus term manually created metadata fields. We used InQuery (version 3.1p1) from the University of Massachusetts for this purpose.[2]

Selecting top n words from top m retrieved documents. Optimal values of n and m depend on the nature of the side collection and the test collection, and in particular on the "closeness" between them. These factors are difficult to characterize without experimentation, so we tried the top 10, 20, 50, and 100 documents, and, for each, the top 10, 20, 30, 40, and 50 words (see Table 2 and 3). Terms are ranked by their cumulative Okapi weight

[2] This was the only stage in our experiments in which retrieval speed was a significant factor. In future experiments we plan to use Zettair or Indri for this step, thus allowing us to explore a broader range of parameter settings.

among the top n documents with a restriction that a selected word should appear in at least 3 of the top n documents (this restriction was intended to prevent pathological cases from dominating the results). The selected words were finally concatenated with the original ASR text to form a expanded segment that was then available for indexing.

We repeated the entire process for each of the 8,104 segments. With several variants of expanded document collections generated in this way and the original document collection, we were able to use the same set of queries to run a set of directly comparable ranked retrieval experiments. Retrieval results were then compared so that we could compare the relative effectiveness of each parameter setting.

2.2 Document expansion using thesaurus relationships

One unique feature of the CLEF-2005 CL-SR test collection is the availability of computationally tractable thesaurus relationships. Thesaurus terms were manually assigned to each segment by subject matter experts at the Survivors of the Shoah Visual History Foundation, thesaurus terms that were automatically assigned by two k-Nearest Neighbor (kNN) classifiers based solely on ASR results are distributed with the test collection, and a Perl script is provided to expand the indexable term set for any of those three sets of thesaurus terms using synonym, part-whole, or is-a relationships. We know from our 2004 experiments that indexing manually assigned thesaurus terms can yield retrieval effectiveness that is substantially higher than that obtained by indexing ASR text, but at the time we ran those experiments we did not have easy access to the synonym, part-whole, or is-a relationships. Preliminary experiments with indexing synonyms and the two types of broader terms (part-whole and is-a) indicated that synonyms appeared to be the most promising for the training set. We therefore chose to focus on the synonyms relationship for our CLEF-2005 experiments.

Document expansion using thesaurus synonyms was easily implemented for any condition that contained manually or automatically assigned thesaurus terms. In our 2004 experiments, we found that concatenating manually created summaries and manually assigned thesaurus terms yielded better results than indexing either alone. We therefore used those two fields plus manually assigned person names (which on their own are of little use) as the baseline case for our synonym expansion experiments. For the expanded condition, we further added all known synonyms for each manually assigned thesaurus terms (we did this by using the Perl script to generate a collection with the synonyms included). A similar design could be tried with automatically assigned thesaurus terms, but our initial experiments with the training set showed no gains when synonym expansion was applied to automatically assigned keywords (concatenated with ASR text), so we did not pursue that option further for our CLEF-2005 experiments.

We realized after the collection was released (and, indeed, after our CLEF experiments were completed) that referring to these expansion terms a "synonyms" was inaccurate. For example, the thesaurus term :extended family members" would be expanded with "aunts," "cousins," "great-grandparents," "sisters-in-law," "brothers-in-law," "daughters-in-law," "fathers-in-law," and "mothers-in-law." Although these terms could indeed serve as a useful basis for document expansion, they are not synonyms. Rather, the thesaurus relationship that is being expressed is "use for" (i.e., instructing an indexer or a searcher to use "extended family members" for "cousins"). We have continued to refer to this relationship as "synonym" in this paper in order to remain consistent with the terminology used in the documentation accompanying the CLEF-2005 test collection, but the reader should bear this distinction in mind when interpreting our results.

2.3 Query expansion using blind relevance feedback

Our first document expansion technique, based on identifying useful terms using an initial search, is generally known as "blind relevance feedback" or "pseudo-relevance feedback" because it is computationally similar to techniques that were originally developed to select expansion terms from documents that a searcher had indicated to be relevant. For blind relevance feedback, we act as if the top-ranked documents were relevant (i.e., our decisions are made blindly with respect to their actual relevance). Using blind relevance feedback from the collection being searched as a basis for query expansion has been shown to work well when the test collection is very large (thus increasing the likelihood that some top-ranked documents will actually be relevant) and when the collection contains text generated through a process with few errors (e.g., professionally edited newswire stories), thus increasing the likelihood that useful expansion terms can be reliably identified).

Unfortunately, the CLEF-2005 CL-SR test collection satisfies neither condition. We nonetheless performed query expansion using the collection to be searched rather than using the available side collection because that provided a cleaner design for exploring the interaction between query and document expansion. When both expansion techniques were applied, we ran document expansion first, and then used the resulting collection as a basis for query expansion. We tried the top 5, 10, 15, and 20 Okapi words respectively from top 10, 20, or 30 top documents using the training topics and found that top 5 words from top 20 documents gave us the best results. We also tried limiting the our choice of top words to those that appeared in at least 1, 2, or 3 of the top m documents. We found that 2 was the best choice for this parameter on the training topics. Those parameters (top 5 words appearing in at least 2 of the top 20 documents) were therefore used for query expansion in all of our official submissions. The effectiveness of blind relevance feedback for query expansion is known to depend on the difficulty of the original query [11]. We did not, however, adjust our query expansion parameters based on query length or other factors for the experiments reported in this paper.

2.4 Cross-language retrieval using statistical translation

Cross-language speech retrieval has previously been explored in the context of broadcast news in the Topic Detection and Tracking Evaluations and in the CLEF-2003 and 2004 CL-SDR evaluations. The usual approach has been to combine monolingual speech retrieval techniques with cross-language text retrieval techniques. That is, the spoken documents are first transcribed into text with an ASR engine and then either the transcribed documents or the query is translated into the other language. Translation can be done using hand-crafted bilingual dictionaries, translation knowledge learned from parallel corpus, or a full-fledged machine translation (MT) systems. Experiments with newswire text have generally indicated that translation statistics learned form parallel text can be remarkably useful, routinely achieving mean average precision measures close to that which can be achieved with in a comparable monolingual experiment design. Corpus-based translation techniques are, however, sensitive to the degree of topical alignment between the corpus from which the translation statistics are learned and the test collection on which the resulting cross-language retrieval system will be evaluated. The CLEF-2005 CL-SR test collection provides an excellent opportunity to begin to characterize this effect because the topical coverage of that collection is quite different from the topical coverage of the large collections of parallel text that have been assembled for use in other tasks.

The basic idea behind the corpus analysis techniques developed for statistical machine translation is to learn the probability that each term in one language will be translated into each term in another language by counting term co-occurrence in a sentence-aligned parallel corpus. When the training corpus is large enough, this process can produce a reasonably accurate set of translation alternatives, together with a smoothed estimate of the probability that each such mapping will occur. For our experiments, used French topics to search the English speech segments. We used the freely available Giza++ toolkit [6][3] to train translation models with the Europarl parallel corpus [4]. Europarl contains 677,913 automatically aligned sentence pairs in English and French from the European Parliament. We stripped accents from every character and filtered out implausible sentence alignments by eliminating sentence pairs that had a token ratio either smaller than 0.2 or larger than 5; that resulted in 672,247 sentence pairs that were actually used. We started with 10 IBM Model 1 iterations, followed by 5 Hidden Markov Model (HMM) iterations,

[3] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

and ending with 5 IBM Model 4 iterations. The result is a a three-column table that specifies, for each French-English word pair, the normalized translation probability of the English word given the French word.

Unlike dictionary-based techniques, statistical analysis of parallel corpora can yield a potentially infinite set of translation mappings with progressively smaller translation probabilities. Threshold selection to limit the options to the most plausible translations is therefore important. We are presently exploring that question in other work (in the context of news retrieval), but for our CLEF-2005 experiments we simply used the most likely translation as the basis for word-by word query translation from French into English. Preliminary experiments on the training set using probabilistic structured queries with additional translation options did not yield promising results, but time pressure precluded analysis of the reasons for that result. Our CL-SR results should therefore be considered to be an initial baseline from which we would hope to make further improvements.

3 Experiment results

The required run in the CLEF-2005 CL-SR track called for use of the *title* and *description* fields as a basis for formulating queries. We therefore used all words from those fields as the query (a condition we call "TD") for our five official submissions. Stopwords in each query (as well as in each document) were automatically removed by InQuery, which is the retrieval engine that we used for all of our experiments. Stemming of the queries and documents was performed automatically by InQuery using kstem. Term-by-term failure analysis is more tractable for shorter queries, so we also ran one set of experiments (described in Section 4) using only terms from the title field of the topic description (a condition we call "T"). Statistical significance is reported for $p < 0.05$ by a Wilcoxon signed rank test for paired samples. Recently reported results from resampling TREC results report 85% confidence for observed differences larger than 10% (relative) at $p < 0.05$ when 25 topics are used (and 90% confidence for 20% relative differences under the same conditions) [8].[4]

3.1 Official evaluation results

Table 1 shows the experiment conditions and the mean average precision for the five official runs that we submitted. Not surprisingly, the two runs with manual metadata (PIQ person names, manual keywords and their thesaurus synonyms, and segment summary) yielded the best results. Comparing the first two columns reveals that document expansion was indeed helpful (see Section 3.2 for more details on this). Enriching the ASR text with automatically generated keywords (i.e., comparing asr.en.qe with autokey+asr.en.qe) produced a similar beneficial effect.[5] This is consistent with the results we obtained with the training set, in which ASR alone yielded a mean average precision of 0.055, AUTOKEYWORD2004A2 alone produced 0.032, and combining both in a single index yielded 0.066. Comparing the last two columns, CL-SR using one-best translation with synonym-expanded metadata achieved about 79% of monolingual effectiveness under similar conditions. This is typical of results seen with one-best query translation in other settings.

3.2 Document expansion results

Table 2 and 3 show unofficial results for experiments with document expansion on the training and evaluation sets respectively. Three parameters were varied: (1) the number of words from each segment used to formulate the expansion query, (2) the number of top-ranked documents from which expansion words were selected, and (3) the number of expansion words that were selected. All parameter settings produced improvements over the no-expansion condition for both

[4] Sanderson and Zobel report the best results from a paired t-test, but the Wilcoxon was reported to be nearly as sensitive.

[5] For all the experiments reported in this paper that involve ASR text, we used the ASR text in ASRTEXT2004A.

run name	asr.en.qe	asr.de.en.qe	autokey+asr.en.qe	metadata+syn.fr2en.qe	metadata+syn.en.qe
CL-SR?	monolingual	monolingual	monolingual	CL-SR	monolingual
doc fields	ASR text	ASR text	ASR text auto-keyword	metadata synonym	metadata synonym
doc exp?	×	√	×	×	×
syn exp?	×	×	×	√	√
MAP	0.1102	0.1275	0.1288	0.2476	0.3129

Table 1: Mean average precision (MAP) for official runs, TD queries with automatic query expansion. ASR text: ASRTEXT2004A; auto-keyword: AUTOKEYWORD2004A2; metadata: NAME, MANUALKEYWORD, and SUMMARY; synonym: thesaurus synonyms of MANUALKEYWORD.

the training and evaluation sets. In the training condition, 40-word expansion queries and selection of the 20 most selective words from the top 50 documents yielded the best retrieval effectiveness (the bolded value), so that condition was used in our official asr.de.en.qe submission. This yielded a 6% apparent relative improvement over the unexpanded condition on the evaluation collection that was not statistically significant, far smaller than the 24% statistically significant relative improvement observed on the training collection. Exploration of the parameter space on the evaluation collection indicated that the optimal parameter setting would have yielded less than a 9% relative improvement over the unexpanded condition. This substantial difference between the training and evaluation sets suggests that the utility of document expansion is somewhat variable, and that topic-specific tuning might be productive.

	formulating query with top 20 words					formulating query with top 40 words				
$m \setminus n$	10	20	30	40	50	10	20	30	40	50
10	0.0582	0.0605	0.0602	0.0604	0.0609	0.0630	0.0588	0.0587	0.0596	0.0592
20	0.0607	0.0597	0.0600	0.0569	0.0569	0.0600	0.0592	0.0598	0.0596	0.0600
50	0.0612	0.0601	0.0601	0.0594	0.0569	0.0615	**0.0681**	0.0613	0.0595	0.0580
100	0.0623	0.0641	0.0616	0.0604	0.0604	0.0614	0.0628	0.0622	0.0619	0.0630
baseline (without document expansion): 0.0551										

Table 2: Monolingual retrieval MAP with document expansion. TD queries, 38 training topics. m: the number of top documents used. n: the number of top words selected from top m documents based on Okapi weight.

formulating query with top 40 words					
$m \setminus n$	10	20	30	40	50
10	0.0995	0.0993	0.1004	0.1007	0.1030
20	0.1060	0.1005	0.1055	**0.1072**	0.1063
50	0.1041	0.1048	0.1040	0.1017	0.1048
100	0.1018	0.1010	0.1024	0.1042	0.1029
baseline (without document expansion): 0.0987					

Table 3: Monolingual retrieval MAP with document expansion. TD queries, 25 test topics. m: the number of top documents used. n: the number of top words selected from top m documents based on Okapi weight.

3.3 Query expansion results

Remarkably, query expansion using the same small collection appeared to be helpful under every condition that we tried (see Table 4), although the observed increases in mean average precision were statistically significant only for two of the five conditions (asr.de.fr2en and autokey+asr). Remarkably, the relative and absolute increases in mean average precision were larger when searching ASR text than when searching metadata. The table shows results on the evaluation topics for the the best parameter settings that were learned using only the training topics (top 5 words from top 20 retrieved segments). Table 5 illustrates the sensitivity of mean average precision to those parameter settings on the training set.

	asr.de.en	asr.de.fr2en	autokey+asr	metadata+syn	metadata+syn.fr2en
Unexpanded	0.1048	0.0814	0.1113	0.3011	0.2327
Query Expansion	0.1275	0.1178	0.1288	0.3129	0.2476

Table 4: Query expansion using blind relevance feedback helps speech retrieval, TD queries, 25 test topics, top 5 words from top 20 retrieved documents.

$m \setminus n$	5	10	15	20
10	0.3128	0.3064	0.3036	0.3069
20	**0.3170**	0.3115	0.3057	0.3102
30	0.2949	0.3136	0.3078	0.3018
baseline (without expansion): 0.3011				

Table 5: Query expansion using blind relevance feedback, TD queries, 38 training topics, top n words from top m retrieved documents.

3.4 Synonym expansion results

As Table 6 shows, expanding manually assigned thesaurus terms with synonyms yielded a 4% apparent improvement that was not statistically significant on the training set and a 3% apparent relative reduction in mean average precision that was not statistically significant on the evaluation set. This somewhat surprising result may reflect a bias in the vocabulary used in the topic descriptions that favors the more “proper” terminology that was designated as the preferred expression for a thesaurus entry.

	38 training topics	25 test topics
no expansion	0.2740	0.3090
synonym expansion	0.2848	0.3011

Table 6: Synonym expansion for manual keywords, TD queries

4 Failure analysis

Our best fully automatic official run (autokey+asr.en.qe) yielded just 41% of the mean average precision achieved by our best official run using manual metadata (metadata+syn.en.de). The mean across topics masks quite a lot of variation, however. Understanding the causes for that variation is important if we are to target future system development efforts productively.

One way to establish an appropriate upper bound for the retrieval effectiveness of an ASR-based system would be to manually transcribe the entire collection. Doing so would be expensive and time consuming, however (requiring perhaps 50 person-months of effort). Manually produced metadata, which already exists, offers an alternative baseline. Query-by-query analysis can be done with queries of any length, but shorter queries offer the additional potential for term-by-term analysis. We therefore chose to analyze an unofficial run on title-only queries with ASR text alone (i.e., with no document expansion, no query expansion, and no automatically assigned thesaurus terms). In order to maximize the number of available topics, we combined the training and test sets for this analysis.

Figure 1 shows a query-by-query comparison of average precision between ASR and metadata for the 32 topics for which metadata yielded a mean average precision above 0.2. The light gray bars at the bottom show the average precision achieved for each topic using ASR, while the darker bars above show how much better metadata did. For example, topic 1179 yielded an average precision of 0.53 with ASR and an average precision of 0.78 with metadata (i.e., the darker bar is the difference). We chose to focus on those 32 topics because the 31 cases in which metadata yielded poor results offered little scope for comparison (i.e., ASR also yielded poor results in every one of those cases).[6] After removing stopwords from each of the remaining 32 title queries, we counted the total number of segments that contained a stemmed match for each query word in the ASR text and in the metadata.

As Table 7 shows, in every case in which a query word was completely absent from all 8,104 ASR segments resulted in very poor retrieval effectiveness for the ASR conditions. Interestingly, all of the eight missing words are proper names. A similar pattern is evident to a lesser extent for the other four queries that performed similarly poorly with ASR, with “sobibor,” “minsk,” “wallenberg,” and “female,” appearing far less in ASR than in metadata. Three of the terms (“female,” “sinti,” and “roma” might result from use of other words (e.g., “women” or “gypsies”) in the spoken content, but the other seven cases suggest that name recognition accuracy by the ASR system may be a contributing factor. Interestingly, similar problems are not evident for many other proper names (e.g., “bulgaria,” “shanghai,” “italy,” and “sweden”). One obvious difference between these proper names and the names that were missed by ASR is that the “easier” names are common in many types of current documents, while the “harder” ones tend to be more specific to the Holocaust. Modeling the usage of frequently used names is easier for the developers of ASR systems because much more training data is available, and that may account for at least some of the difference. This suggests that domain-tuned techniques for language modeling with the ASR system and/or domain-adapted techniques for accommodating weaknesses in the ASR language model might be a productive line of investigation.

For the rest of 22 queries listed in the table, query word coverage by both ASR and metadata are quite comparable to each other. For space limitation, we don’t list the actual numbers in the table.

5 Conclusion

This year’s CLEF CL-SR track has provided an excellent opportunity to study the problem of speech retrieval in a domain other than broadcast news. The availability of document fields from different sources – some with manually created metadata, some with automatically generated text and keywords – made it possible to explore a variety of contrastive conditions and data fusion techniques. The availability of a large side collection also provided an opportunity to re-examine the potential of document expansion to mitigate the effect of recognition errors. Through a series of experiments with the 38 training topics and the 25 test topics, we were able to show that a combination of document expansion using a side collection and query expansion using the collection being searched could improve speech retrieval effectiveness and that tuning the expansion parameters on a set of 38 training topics yielded near-optimal improvements on the 25 evaluation topics (when searching the same collection). A query-by-query analysis of query term coverage

[6] In future work, we will want to look at what makes those 31 topics hard.

topic ID	query words	in both ASR and metadata	only in metadata
1188	volkswagen jews	jews	volkswagen
1630	eichmann witnesses	witness	eichmann
2400	sobibor death camp	sobibor(5/13) death camp	
2185	sinti roma holocaust	holocaust	sinti roma
1628	slave labor aeg telefunken	slave labor	telefunken
1187	ig farben labor camps	labor camps	ig farben
1337	art auschwitz	art auschwitz	
1446	minsk ghetto underground	minsk(21/71) ghetto underground	
2264	abusive female personnel	abusive female(8/81) personnel	
1330	wallenberg eichmann	wallenberg(3/16)	eichmann
1850	wallenberg rescue jews	wallenberg(3/16) rescue jews	
1414	fort ontario refugee camp	fort ontario refugee camp	
1620	jewish nurses concentration camps	jewish nurses concentration camps	
2367	war crime trial participants	war crime trial participants	
2232	french internment camps	french internment camps	
2000	post liberation experience	post liberation experience	
14313	birkenau daily life	birkenau daily life	
2198	sonderkommando	sonderkommando	
1829	jewish gentile relations poland	jewish gentile relations poland	
1181	sonderkommando auschwitz	sonderkommando auschwitz	
1225	liberation buchenwald dachau	liberation buchenwald dachau	
14312	jewish kapos	jewish kapos	
1192	kindertransport	kindertransport	
2404	decision migration australia	decision migration australia	
2055	flight denmark sweden	flight denmark sweden	
1871	dp camps american zone	dp camps american zone	
1427	rescue danish children	rescue danish children	
2213	persecution jews italy	persecution jews italy	
1877	kindertransport possessions	kindertransport possessions	
2384	red cross holocaust	red cross holocaust	
1605	jews shanghai	jews shanghai	
1179	bulgaria saved jews	bulgaria saved jews	

Table 7: Query term coverage between ASR text and metadata. In parentheses are (number of segments with ASR text containing the word / number of segments with metadata containing the word). Listed in the same order as Figure 1 (most difficult for ASR at the top).

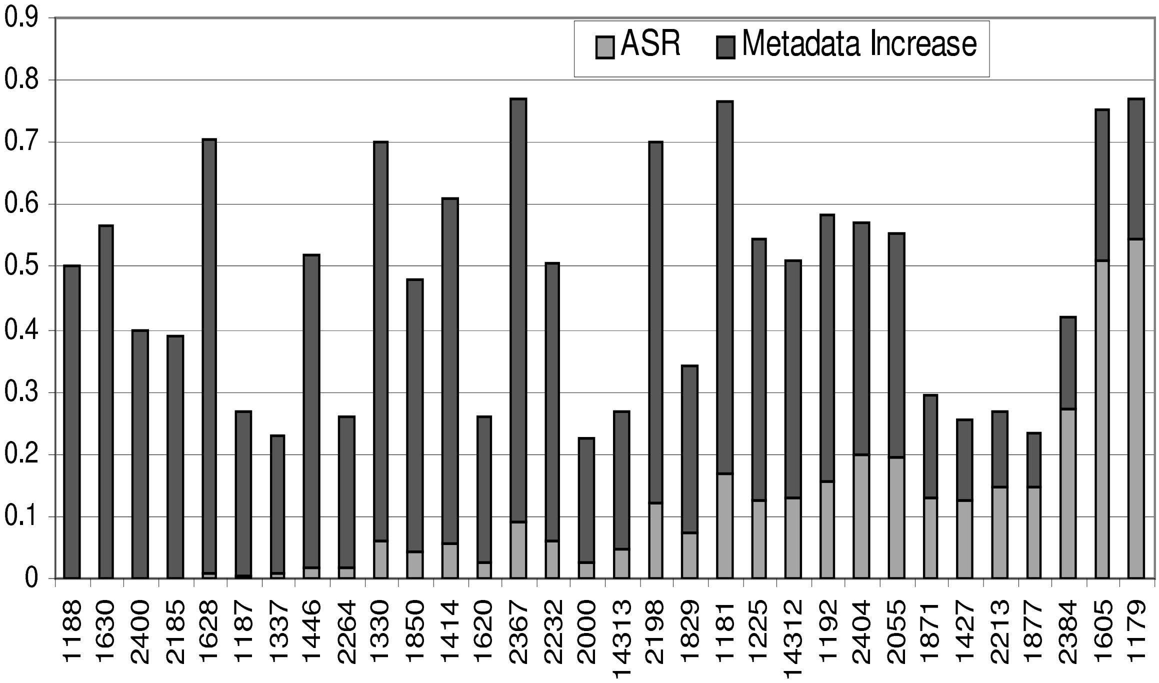

Figure 1: Query-by-query comparison of retrieval effectiveness (average precision) between ASR text and metadata, 32 title queries with average precision of metadata equal to or higher than 0.2. in increasing order of (ASR MAP) / (metadata MAP).

revealed that failure to reliably recognize domain-specific named entities was a possible cause for a substantial number of the cases in which very poor results were observed from ASR-based searches.

Three areas for future work are clearly indicated by the results that we have obtained. First, we are looking forward to working with ASR systems with improved accuracy and with word lattices that may allow us to partially mitigate the effect of recognition errors in the one-best transcription. Second, we are interested in extending our baseline cross-language speech retrieval results to explore techniques that accommodate both translation and recognition uncertainty. And finally, we hope to explore a broader range of document expansion techniques that include parameter settings that are adapted to observable document characteristics (e.g., length or clarity measures) and sequence-based expansion (e.g., selectively importing location names from earlier segments). We're therefore looking forward to the CLEF-2006 CL-SR track!

Acknowledgments

The authors would like to thank the MALACH project teams at the University of Maryland, IBM and the Survivors of the Shoah Visual History foundation for creating the test collection, and Okan Kolak and Adam Lopez for their help with using Giza++. This work has been supported in part by NSF IIS award 0122466 (MALACH). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] James Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer-Verlag London, UK, 2001.

[2] Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–344. ACM Press, July 2003.

[3] John S. Garofolo, Cedric G. P. Auzanne, and Ellen E. Voorhees. The TREC spoken document retrieval track: A successful story. In *Proceedings of the Nineth Text REtrieval Conference (TREC-9)*, 2000. http://trec.nist.dov.

[4] Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft. 2002.

[5] Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–38, 2004.

[6] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL'00*, pages 440–447, Hongkong, China, October 2000.

[7] S. E. Robertson and Karen Sparck-Jones. Simple proven approaches to text retrieval. Cambridge University Computer Laboratory, 1997.

[8] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity and reliability. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, August 2005.

[9] Amit Singhal, John Choi, Donald Hindle, and Fernado Pereira. ATT at TREC-7. In *The Seventh Text REtrieval Conference*, pages 239–252, November 1998. http://trec.nist.gov.

[10] Amit Singual and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41. ACM Press, August 1999.

[11] Ellen M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 69–77, 2003. http://trec.nist.gov.

[12] Jinxi Xu and Ralph Weischedel. TREC-9 cross-lingual retrieval at BBN. In *The Nineth Text REtrieval Conference*, November 2000. http://trec.nist.gov.